**INFO290: Data Mining & Analytics  
Clustering lab / homework**

**Required Tools:**

* A k-means clustering implementation (scikit-learn / R / matlab / etc)
* A silhouette coefficient function ( [scikit-learn](http://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html) and [matlab](http://www.mathworks.com/help/stats/k-means-clustering.html?refresh=true#brah7fp-1) have these / or code your own)

**Materials: Please document your responses on the google spreadsheets below (assign yourself to a group by writing your CalNet usernamesnext to a group number)**

[**Groups 1-10**](https://docs.google.com/spreadsheets/d/1UwKjjNUK-eNkcoQ7r86syzyJ3nQwBd0T4sRJ6XtfIA4/edit?usp=sharing)

[**Groups 11-20**](https://docs.google.com/spreadsheets/d/1-BRmbSt_TrjjvcNh1fCe2cmmglvUzrdP3N08X7dQvYo/edit?usp=sharing)

[**Groups 21-30**](https://docs.google.com/spreadsheets/d/1Wt5qvKD9VZOUdjIemDTGxlVb-WBV66RY-912qINFlxc/edit?usp=sharing)

**In the column that asks for the names of your group members, please use your calnet usernames.**

**Description:**

In class on Tuesday we learned how the k-means clustering algorithm works and metrics that can be used to score the goodness of a clustering. While you could perhaps visually judge the “goodness“ of a cluster given 2D data, as the dimensionality of your data increases, metrics are needed to guide the evaluation.

In this Lab you will perform a series of cluster analyses on the [*yelp\_reviewers.csv* dataset](https://drive.google.com/a/berkeley.edu/file/d/0BycRP8EC1WVVRm9xNHE4Vy14Ykk/view?usp=sharing) which has been newly created from the Lab 2 features (questions 3 to 15 plus some submitted question 16 features).

The objectives of the Lab are to

a) gain experience in running k-means yourself

b) quantitatively evaluating the goodness of a clustering

c) investigate the answer to your own question about this dataset using clustering techniques.

**Note:**

1. To make it easier to match the Q16 ‘special’ feature columns that students created with their descriptions, there is a [**Lab2\_header\_lookups.csv**](https://drive.google.com/a/berkeley.edu/file/d/0BycRP8EC1WVVUmo5d3kwTWlDYlk/view?usp=sharing)file which matches the Q16. feature columns in the dataset with their description.
2. When dealing with features that contain -Inf values or other oddities, consider replacing (imputing) those values with the mean of the column or completely removing reviewers with -Inf values in any of their features. Because you will be using custom features generated from peers from the previous lab, the data quality of all the features can’t be anticipated beforehand.
3. Be sure to read the FAQ at the bottom (particularly if clustering is taking more than 5m)

**Grading**

This Lab is due at 1:59pm next Thursday and will count as a homework. Grading will be based on completion of questions 1-7 (according to google sheet), quality of research question (don’t pick something too simplistic), and appropriateness of techniques used to investigate question 8.

MAX GROUP SIZE: 2

**Questions/Objectives**

1. Choose an implementation of k-means and specify it in the google sheet
2. What is the best choice of *k* according to the silhouette metric for clustering q4-q6 (# of cool, funny, useful votes) combined. Only consider ***2*** <= *k* <= 8.
   1. **NOTE:** You may use the max, as you did in question 2. For a more in-depth understanding of cluster analysis with silhouette, look [ [here](http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html) ]
3. Answer question 2 but using the log of the features (q**8**-q10)
4. Answer question 2 but using the percentage of the features (q11-q13)
5. Inspect the [best] clustering generated from question 4
   1. List the number of data points in each cluster (eg. C1: 2,000 C2: 4,200 etc)
   2. Were there clusters that represented very funny but useless reviewers?
   3. How many reviewers were in the cluster that represented relatively equal strength in all voting categories (assuming such a cluster exists in your clustering)?
6. Cluster the dataset using *k = 5* and using features q**8**-q13 (log and % type votes) and q14 (most active year feature) and the natural log of q15 (avg review chars.)
   1. What is the silhouette metric for this clustering?  
      You may use the max, as you did in question 2. For a more in-depth understanding of cluster analysis with silhouette, look [ [here](http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html) ]
   2. What was the average “number of reviews per reviewer (q3)” among the points in each of the clusters (eg. C1: 1.4 C2: 4.2 C3: 3.4 etc)
7. Cluster the dataset using the features described in question 6 + every group’s question 16 features (you may drop features with high incidents of -Inf / blank / or NaN values). It is suggested that you perform some form of normalization on these question 16 features so as not to over bias the clustering towards the larger magnitude features.
   1. Using the silhouette metric, what was the best *k*?
   2. Using the the sum of within cluster variance metric with the elbow method what was the best k?
8. For this question please come up with your own question about this dataset and using a clustering technique as part of your method of answering it. This question answer should be submitted to a bCourses assignment in the form a pdf. This report is expected to be between 2 and 3 pages. It is meant to give you practice writing up your results. The report should have the following sections:
   1. RESEARCH QUESTION: Describe what it is you want to ask of this dataset  
       (the question can not be the same as question 9)
   2. DATASET: describe the yelp\_reviews to yelp\_reviewers transformation
      1. subsection called FEATURES SELECTED: briefly describe the features from the dataset you choose (at least 1 features should be from a q16)
   3. METHODS: Briefly describe the methods you used (such as k-means) and why you used them
   4. RESULTS: Here you can tell the story of how you investigated the question (plots are always nice) and the conclusions you drew.

**NOTE:** If you are hypothesizing that clusters might be formed with respect to a particular attribute, consider withholding that attribute from clustering and then looking to see what the mean value or distribution of that attribute is among the formed clusters.

9. Bonus question (+15%) - Reviewer overlap: Create a dataset with f reviewers as the rows and business\_ids as the columns (there are a lot) where the feature value is is ‘1’ if the reviewer has written a review for that business and ‘0’ if not. Use the methods described in this assignment to answer the question of how much overlap of businesses reviewed exists among reviewers in this dataset. Append this answer to your question 8 document.